D.D. Neustroev, D.I. Kurmanova
Ural Federal University named after the first President of Russia B.N. Yeltsin
Yekaterinburg, Russia

INTERPRETABILITY OF MACHINE LEARNING MODELS

Abstract: As machine learning algorithms have been used in many areas of our lives, for example, self-driving cars, healthcare, and the financial industry, the problem of trust is becoming even more urgent. To trust the decisions that the algorithms adopt, we need to understand the nature of their occurrence, so often we need not only a theoretical understanding of their work, but also special tools that would explain the origin of the findings within the algorithms themselves and present the withdrawn information in an informative and accessible form.

This article will list some problems related to the interpretation of machine learning algorithms, as well as the desired properties of interpreted models that can improve the perception of algorithms and increase people's confidence in the decisions made. In the following, some visual analytics tools will be discussed, as well as one of the model-agnostic methods, LIME, which studies the model locally around the prediction and explains any classifier.

Keywords: machine learning, explanation, model interpretability, predictions, transparency.

Д.Д. Неустроев, Д.И. Курманова
Уральский федеральный университет имени первого Президента России Б.Н. Ельцина
Екатеринбург, Россия

ИНТЕРПРЕТИРУЕМОСТЬ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Аннотация: Поскольку алгоритмы машинного обучения стали применяться во многих сферах нашей жизни, например, беспилотные автомобили, здравоохранение, финансовая индустрия, то проблема доверия становится все более актуальной. Чтобы довериться решениям, которые принимают алгоритмы, мы должны понимать природу их возникновения, поэтому часто требуется не только теоретическое понимание их работы, но и специальные инструменты, которые бы объяснили происхождения выводов внутри самих алгоритмов и представили изъятую информацию в информативном и доступном для человека виде.

В этой статье будут перечислены некоторые проблемы, касающиеся интерпретации алгоритмов машинного обучения, а также желаемые свойства интерпретируемых моделей, способных улучшить восприятие алгоритмов и увеличить у человека доверие к принятым решениям. В дальнейшем будет рассказано про некоторые инструменты визуальной аналитики, а также один из модельно-агностических методов – LIME, который изучает модель локально вокруг предсказания и объясняет любой классификатор.

Ключевые слова: машинное обучение, объяснение, интерпретация моделей, предсказания, прозрачность.

In connection with the increasing availability of computing power to machine learning, there is an increasing range of applications in our daily life. Along with the growing amount of information in the information age, algorithms become even more relevant and in demand. The knowledge stored in the data is simply not extracted because of their

multidimensionality and large volume. With proper selection of the model, preprocessing of the input data and some other important steps, the algorithms cope well with the tasks. But the question arises: can we trust the predictions?

Interpretability

Interpretation of machine learning is most often understood as information accessible to humans, which gives a greater degree of understanding of the internal workings of the model. Interpretability of models generates an increase in confidence in the built models from a person who at least does not make mistakes where people do not make mistakes.

There are several basic properties that the interpreted model would preferably possess:

1) **Transparency**. Ideally, transparency should be applied at the level of the learning algorithm to adjust the parameters in real time, but modern deep learning algorithms do not have this degree of transparency. On the other hand, transparency is often hampered by either corporate protection or unreadable code.

2) **Comprehensibility**. The main share of misunderstanding arises because of the complexity of the perception of information at the output or its absence. Associations studied using controlled learning algorithms do not guarantee reflection of cause-effect relationships. But even with complex conversion of functions it may be less clear than the original one.

3) **Informativeness.** If the goal of machine learning is to reduce the number of errors, then the goal of the real world is to obtain useful information for its further use. The degree of informativeness is not quantified because everything depends on the ability to perceive information and the level of development of intuition in a group of people who directly interact with algorithms. But often there are situations when there is not enough information for logical conclusions.

Visual analytics systems

This subsection will be devoted only to two systems of visual analytics:

1) **Prospector** – an interactive visual analytics system that allows for a better assessment of interpretability. The features of this system are interactive diagnostics of partial dependence and local verification of a specific data prediction. Through interactive customization, users can interact with models at a deeper level than with conventional tools.

2) Rivelo. Creating an interpretable model is a laborious task and also reduces performance, and testing the model behavior on a previously known dataset often does not provide a sufficient level of understanding due to the lack of transparency in the operation of the algorithm. A visual analytics interface was created to solve problems – Rivelo, allowing analysts to understand better predictions at the instance level. First, the system generates explanations for each data element and creates a list of functions with which the user further interacts for more detailed study, visual presentation and comparison with other data elements.

LIME

Confidence in predictions by the user is one of the important indicators of the model, otherwise blind faith can lead to disastrous consequences. The developers of the LIME method purposefully stop at this problem, offering a solution in the form of several predictions and explanations, explaining the predictions of any classifier or regressor by local approximation of the interpreted model. The main task of the algorithm is to provide the user with information about when the system is mistaken in order to generate more user confidence and avoid stupid mistakes. Thus, the LIME method generates a qualitative understanding of the interaction of the instance in a textual or visual form.

Model explanations make it possible to make a better choice between models, train unreliable models, increase user confidence, and obtain more detailed information about forecasts. But you need to find a compromise between the complexity of the algorithm and its interpretability. By increasing the quality of the predictive characteristics of the models, the internal structure and complexity increases, so such models are more difficult to understand. Therefore, when using or creating special tools for interpretation, one should always find a certain balance between these characteristics, maximally achieving the main goal - to understand the reason for which a certain decision was made.

REFERENCES

1. Lipton, Zachary C., The Mythos of Model Interpretability. ICML Workshop on Human Interpretability of Machine Learning 2016.

2. Burrell, Jenna. 2016. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. Big Data & Society January-June 2016: 1-12.

3. Paolo Tamagoni, Josua Krause, Aritra Dasgupta, and Enrico Bertini. 2017. Interpreting Black-Box Classifiers Using Instance-Level Visual Explanations. In Proceedings of HILDA’17, Chicago, IL, USA, May 14, 2017, 6 pages.

4. Krause, Josua, Adam Perer and Kenney Ng. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. CHI 2016.

5. Ribeiro, Marco Tulio, Sameer Singh and Carlos Guestrin. «Why Should I Trust You?»: Explaining the Predictions of Any Classifier. HLT-NAACL Demos, 2016.

6. Christopher Molnar, Interpretable Machine Learning. A Guide for Making Black Box Models Explainable // 2018-12-05. [Electronic recource]. URL: https://christophm.github.io/interpretable-ml-book/ (04.12.2018).

7. Ribeiro, Marco Tulio, Sameer Singh and Carlos Guestrin. Introduction to Local Interpretable Model-Agnostic Explanations (LIME). [Электронный ресурс] // Oreilly, August 12, 2016. [Electronic recource]. URL: https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime (05.12.2018).

8. Lilian Weng. How to Explain the Prediction of a Machine Learning Model? // Aug 1, 2017. [Electronic recource]. URL: https://lilianweng.github.io/lil-log/2017/08/01/how-to-explain-the-prediction-of-a-machine-learning-model.html (05.12.2018).